JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 113, D17121, doi:10.1029/2007JD009536, 2008

Interval-based statistical validation of operational seasonal forecasts in Spain conditioned to El Niño–Southern Oscillation events

C. Sordo,[1] M. D. Frías,[1] S. Herrera,[1] A. S. Cofiño,[1] and J. M. Gutiérrez[1]

Received 25 October 2007; revised 13 September 2007; accepted 8 July 2008; published 13 September 2008.

[1] As opposed to the tropics, operational seasonal forecasting systems have shown little or no skill in European midlatitudes. In this paper we explore the potential source of predictability in this region given by El Niño–Southern Oscillation (ENSO) events; in particular we analyze winter rainfall in Spain. First, we apply a simple statistical method to assess the teleconnections between rainfall records in 123 gauges over Spain and ENSO events during the last 40 years. A significant teleconnection for dry winter episodes is found associated with La Niña events, extending the results obtained in previous studies. Then, we adapt the statistical method to perform operational seasonal forecasts validation conditioned to ENSO events; in particular we consider a state-of-the-art operational model, the System2 from ECMWF. The validation method defines a forecast interval to account for the ensemble spread, and applies a simple skill measure based on the proportion of hits (observations falling into the forecast interval) compared with a random forecast. As a result, we uncover the significant skill of operational seasonal predictions for reproducing the dry winter episodes associated with La Niña events (a window of opportunity for operational seasonal forecast in midlatitudes). Finally, the results are improved using statistical downscaling methods and some sensitivity studies are conducted. The analysis presented in this paper can be extended to other regions under the influence of any seasonal predictability-driving factor.

Citation: Sordo, C., M. D. Frías, S. Herrera, A. S. Cofiño, and J. M. Gutiérrez (2008), Interval-based statistical validation of operational seasonal forecasts in Spain conditioned to El Niño–Southern Oscillation events, *J. Geophys. Res.*, *113*, D17121, doi:10.1029/2007JD009536.

1. Introduction

[2] Seasonal forecast is an active area of research due to its potential socioeconomic benefits for a wide range of end users [*Weiss*, 1982; *Challinor et al.*, 2005; *Thompson et al.*, 2006]. Nowadays, most of the major meteorological institutions around the world have developed Ensemble Prediction Systems (EPS) for operational seasonal forecasting based on coupled atmosphere-ocean general circulation models. Some examples are the ECMWF forecasting Systems [*Anderson et al.*, 2003], the NCEP CFS [*Saha et al.*, 2006], the Australian POAMA [*Wang et al.*, 2001], and the recent European EURO-SIP multimodel resulting from the DEMETER project [*Palmer et al.*, 2004]. The ocean is one of the components that gives potential predictability at seasonal time scale, since it has a large heat capacity and slow adjustment times relative to the atmosphere [*Palmer and Anderson*, 1994].

[3] The skill of current seasonal forecast systems seems to be limited to particular areas and periods. For instance, in low-latitude regions the influence of the ENSO phenomenon is recognized as one of the major sources of predictability [*Hastenrath*, 1995; *Palmer et al.*, 2004; *Derome et al.*, 2005]. Some seasonal predictability has also been identified in midlatitude regions (e.g., in North America) associated with ENSO teleconnections [see, e.g., *Quan et al.*, 2006], and also with other sources such as the persistence of the North Pacific Decadal Oscillation [*Gershunov and Cayan*, 2003] or the state of the land surface at the start of a season, in particular the soil moisture content [*Wang and Kumar*, 1998; *Douville*, 2004]. However, in most of the extratropics the signals predicted by operational models are weak (the ensemble covers most of the climatological range diminishing the signal-to-noise ratio) and do not add valuable information over a climatological forecast. For instance, no clear operational seasonal skill has as yet been found in Europe for 2-m temperature or precipitation, although some studies have pointed to plausible sources of predictability associated, for instance, with the Arctic Oscillation (AO) [*Johansson et al.*, 1998], which seems to have some potential seasonal skill derived from operational forecasts in connection with strong ENSO events [*Derome et al.*, 2005]. Other studies have detected weak skill signals for particular European regions and periods, but no statistical

[1]Department of Applied Mathematics and Computer Science, University of Cantabria, Santander, Spain.

0148-0227/08/2007JD009536$09.00

analysis of the results has been provided in order to distinguish real from spurious skill [*Díez et al.*, 2005].

[4] The main motivation for the present work is the need of a simple and intuitive statistical inference framework able to estimate the skill of the operational forecasting systems for different regions and periods, and also to assess the statistical significance of the results in order to identify spurious skill that may have occurred by chance. This information is demanded by end users from different socioeconomic sectors, since the potential economic benefits of seasonal predictions lie in planning the future (e.g., taking protection measures) when an event is forecasted in a certain region with a prescribed confidence. From this point of view, seasonal predictions should be only provided for those periods where the influence of some predictability-driving factor (e.g., the ENSO phenomenon considered in this study) gives a ''window of opportunity'' with proven skill. However, this requires extending the standard validation, based on temporal-averaged skill measures, to the case of conditional validation. In this paper we explore this problem and provide a simple solution for this conditional validation task. We show that the windows of opportunity can be objectively found in relation to the spread/uncertainty of the forecast.

[5] A common approach used to evaluate the skill of probabilistic binary predictions (e.g., above or below a given threshold) is based on the Relative Operating Characteristics (ROC) curve or the associated economic value (see *Jolliffe and Stephenson* [2003] for more details). However, these measures do not provide a simple and intuitive inference framework suitable to compare the significance of the results against, for instance, a random prediction. Therefore, it would be worthwhile to consider some simple method which could be suitable for this task. Recently, a simple validation approach has been proposed by *Weisheimer et al.* [2005]; they define the bounding box given by the minimum and maximum values of the ensemble members as a prediction interval, thus accounting for the spread of the ensemble. Then, the observed event may fall inside (hit, or correct forecast) or outside (miss) the interval, hence providing a sound and intuitive prediction/validation method for ensemble prediction systems. Moreover, they analyze the relationship between the spread of intervals and the hit rate of the associated forecasts for 2-m temperature; accurate predictions (defined by a fraction larger than 95% of hits with spread smaller than the climatologic range) are found over land areas, in particular over almost 50% of Europe, thus proving the utility of this methodology.

[6] In this paper we further explore the above idea of interval-based prediction considering some improvements of the above method to achieve robust unbiased predictions and introducing a simple inference framework to estimate the significance of the resulting skill. On the one hand, the interval based on extremes is not a robust estimation of the spread; moreover, for variables such as precipitation in midlatitudes this interval usually covers the whole climatological range. To avoid this problem, we define the forecast interval using the interquartile range of the member values, thus providing a robust estimation of the spread. Moreover, as we will show later, this amplitude is appropriate for both low- and mid-latitude predictions. Then, systematic errors are removed by using order statistics (see *Balakrishnan and Cohen* [1991] for an introduction of order statistics). This means that the forecast quantity is the percentile corresponding to the precipitation value, rather than the precipitation per se. Therefore, the prediction interval is defined by the interquartile range of the percentile member values. In this study, observation and model climatologies are divided respectively into quintiles (very wet, wet, normal, dry and very dry), although other percentile values such as deciles/terciles could be used instead to increase/decrease the resolution of the predictions to be validated. Thus, for instance, if the forecast is wet (according to the model climatology) and the observation is also wet (according to the observation climatology) then the prediction can be considered correct; otherwise it can be considered wrong. Note that in this case the prediction would be also correct if the prediction interval were [1, 3], covering from very wet to normal forecasts. This simple validation method allows us to compare the forecasted and observed relative signals (or fluctuations) using simple and intuitive statistical indices (e.g., the frequency of hits). On the other hand, standard statistical inference methods can be used to estimate the significance of the results by comparing the forecast hit rate with that expected from a random prediction.

[7] The study is focused on assessing operational seasonal precipitation predictability in boreal winter DJF (December, January and February) in Spain. As an illustrative example, the statistical method is applied to one state-of-the-art seasonal prediction models, the ECMWF System2 forecast system [*Anderson et al.*, 2003]. A validation conditioned to certain windows of opportunity (for instance associated with ENSO events) is performed. The novel approach is first validated over a tropical region (Peru) with a known seasonal predictability given by the influence of the ENSO phenomenon. A previous study by *Gutiérrez et al.* [2005] shows the possibility of forecasting heavy precipitation episodes some months in advance in Peru associated with strong El Niño events. Here, the interval-based method can add information taking into account the model uncertainty. In Spain, the influence of ENSO events is lower than in tropical latitudes; however, *Pozo-Vázquez et al.* [2005] found a significant teleconnection between La Niña events and negative anomalies in southern Europe in winter, especially in the southwest of the Iberian peninsula. In our work, the study of the teleconnections with ENSO events is extended by using a bigger number of stations, 123 gauges. The statistical method is then applied over Spain to detect possible predictability at seasonal timescales linked to the ENSO phenomenon. This analysis also provides an estimation of the uncertainty associated with the forecasts. The skill of System2 seasonal precipitation forecasts is analyzed by means of the direct model outputs and also in combination with a statistical downscaling method based on analogs. Therefore, the present work also tries to assess the advantages of considering a downscaling method in studies of predictability. Some studies have reported the necessity of this postprocess to extract local skill from the model predictions [*Gutiérrez et al.*, 2005].

[8] The manuscript is structured as follows. In section 2 we describe the data and introduce a new method to analyze the teleconnections with ENSO events applied to Spain. We consider both the direct model outputs and the downscaled values obtained using a statistical method (see section 3).

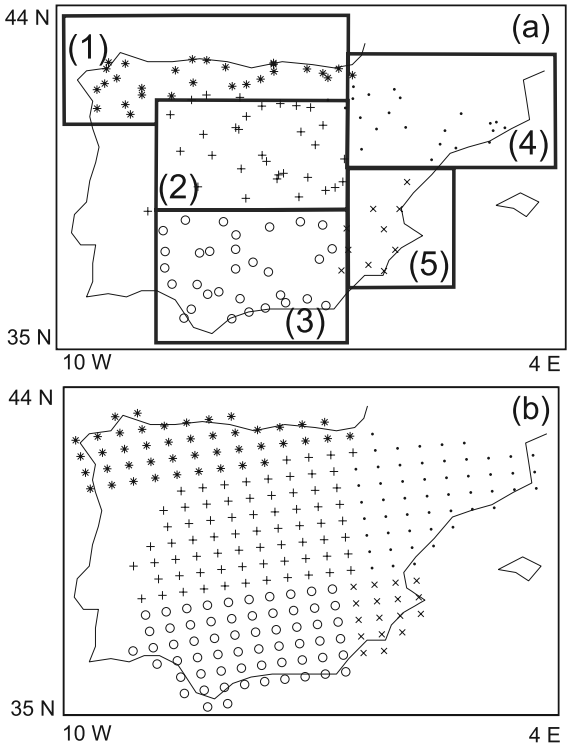

Figure 1. (a) Stations over Spain (123) with precipitation records. The boxes indicate the predictor areas used in the downscaling method over Spain (five regions). (b) Points of the 50 × 50 km Joint Research Centre grid data set. Note that a Lambert azimuthal projection was used in the definition of the grid, so it may not appear regular in the displayed lat-lon projection (interpolated meteorological data source JRC/AGRIFISH Data Base-EC-JRC). Different symbols indicate the five areas considered.

Section 4 describes the validation method. The verification results are shown in section 5. Finally, section 6 shows the results of different sensitivity studies regarding the nature of the observations and the time aggregation period (daily, weekly or monthly). Finally, some conclusions are given in section 7.

2. Data

[9] The seasonal ensemble forecast system selected for this study is System2 from the ECMWF [*Anderson et al.*, 2003]. The atmospheric component has a horizontal resolution TL95 (approximately 1.875° × 1.875° latitude-longitude grid) and 40 levels in the vertical. The ocean model is based on HOPE version 2 and has 29 vertical levels with different resolution from 0.3° in the equator to 1° in higher latitudes. A hindcast covering the period 1987–2004 has been produced for this model from 40 perturbations of initial conditions for integrations initialized on 1 November and 1 May, all running for a 6-month period (see *Anderson et al.* [2003] for more details). In this study we focus on boreal winter season so we analyze the integrations started in November for December–February (DJF) season, characterized by 90 daily values for each of the 40 ensemble members. Note that a maximum of 17 seasons are available for the statistical validation of the model.

[10] The seasonal (DJF) accumulated precipitation values from System2 will be evaluated against observed data.

2.1. Raw Observed Station Data

[11] Station data were provided by the Spanish State Meteorological Agency (AEMET). These stations correspond to 123 gauges covering the period from 1958 to 2004 (see Figure 1a).

[12] According to *Muñoz-Díaz and Rodrigo* [2004], five regions were considered in Spain: north, center, south, north Mediterranean and south Mediterranean, corresponding to

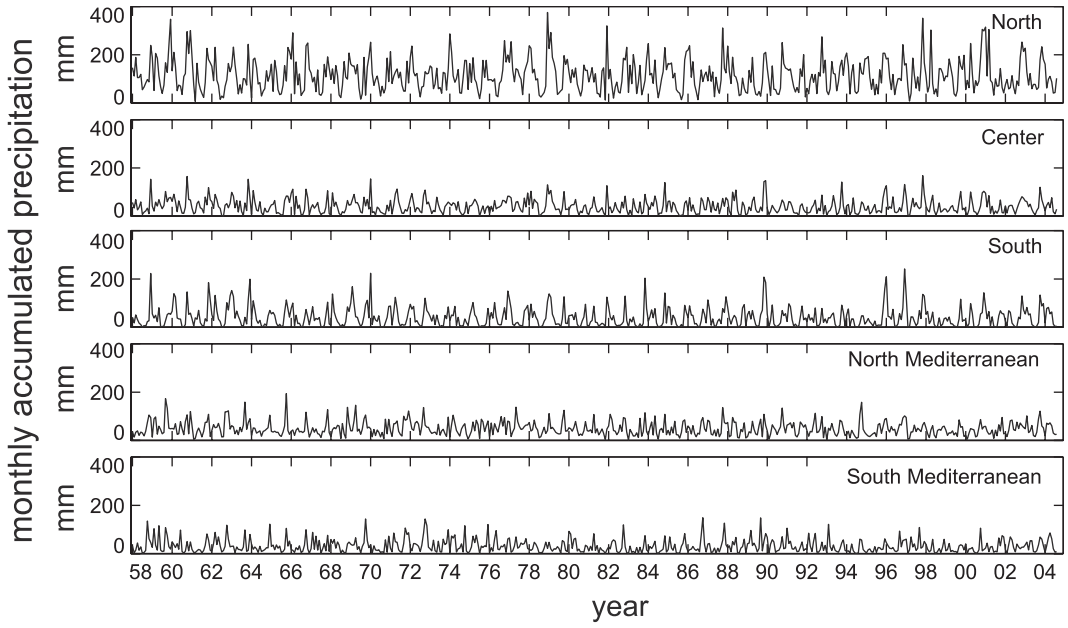

Figure 2. Monthly accumulated precipitation in five regions of Spain.

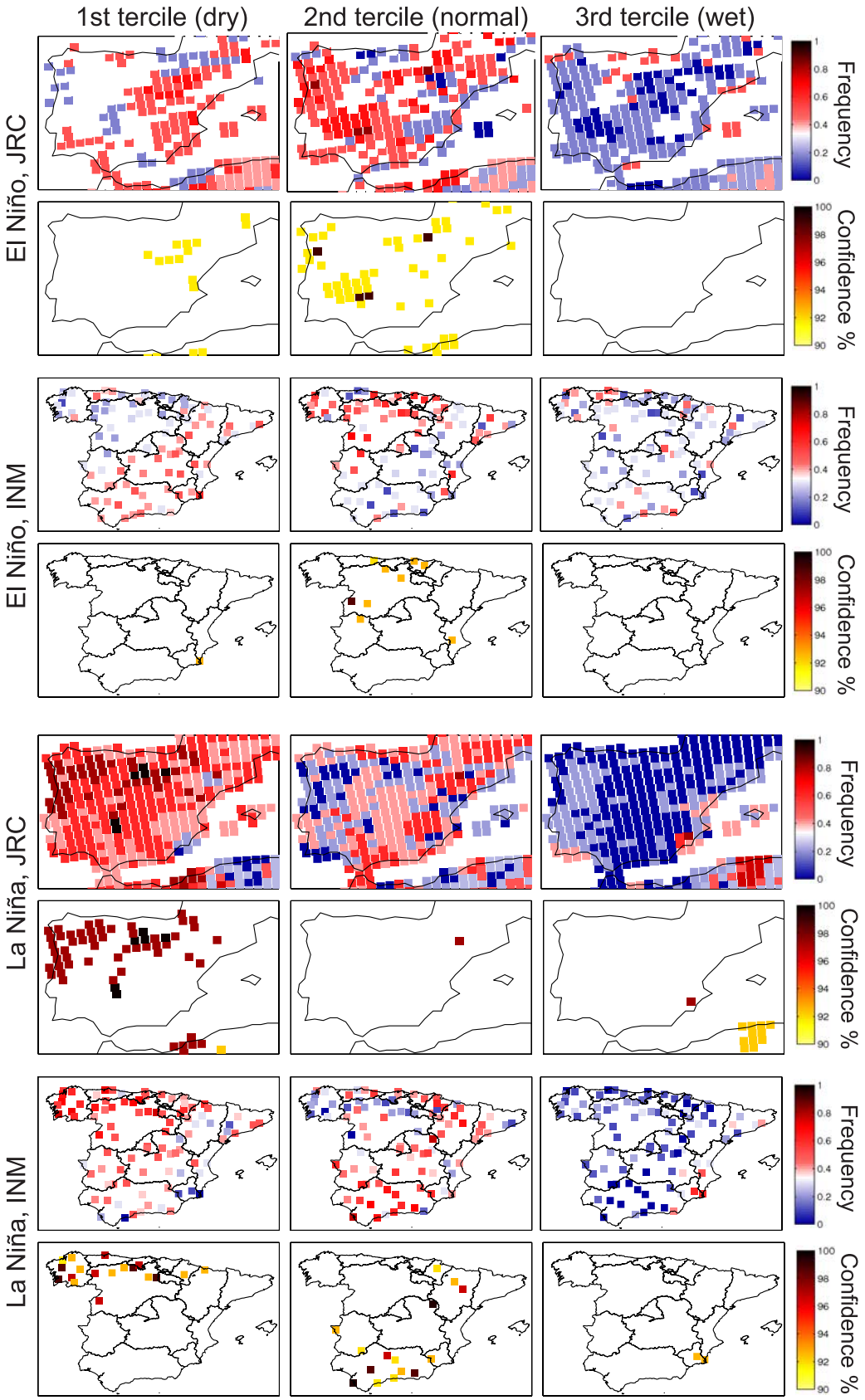

Figure 3. Proportions (relative frequencies) and level of confidence for the dry/normal/wet categories during El Niño and La Niña periods for the JRC gridded observations and for the raw gauge data set from AEMET; note that the climatological proportions are 1/3 in all cases (interpolated meteorological data source JRC/AGRIFISH Data Base-EC-JRC).

labels 1 to 5 in Figure 1a, respectively. Monthly accumulated precipitation time series are represented in Figure 2 showing different precipitation regimes. For instance, the north of Spain shows the highest values of accumulated precipitation (close to 400 mm in some months) whereas the lowest values correspond to the south Mediterranean area. On the other hand, stations in the south show more extreme episodes randomly distributed during the period.

[13] In this paper we present both local and regional results corresponding to the stations' observations and to regional averaged records within the above described homogeneous areas.

2.2. Interpolated Observations

[14] Besides the raw observations, in this paper we also consider interpolated gridded data to check the robustness and consistency of the results. Regional averages are more robust than local observations, but fail to provide spatial detail. Some studies suggest the use of high-resolution gridded data derived from observations as an alternative to raw data for forecast validation [see, e.g., *Osborn and Hulme*, 1997]. Each of the above options has advantages and shortcomings. Gridded data is built using different quality tests to correct inhomogeneities and possible observation errors. Thus, the resulting data set is more homogeneous than the original observations. However, some local information from the original data can be missed during the interpolation process which could modify in some extent the results. For instance, an interpolation of two stations with one of them showing a strong influence of a known teleconnection pattern will lead to unrealistic precipitation values. This could be more appropriate for the validation of general circulation models (GCMs), but do not correspond with the regional behavior of precipitation.

[15] Therefore, besides the raw observations, in this paper we also consider the interpolated MARS-STAT data set from the Joint Research Centre (JRC, http://agrifish.jrc.it), defined on a regular latitude-longitude 50 $\times$ 50 km grid over Spain (see Figure 1b).

2.3. Teleconnection With ENSO Events

[16] In order to assess potential sources of conditional seasonal predictability in Spain we explore the teleconnections of the precipitation series with ENSO events, the main source of skill for seasonal prediction. Several studies have found relationships between the ENSO cycle and different variables in European regions for specific time periods. Focusing on winter precipitation *Pozo-Vázquez et al.* [2005] found a significant teleconnection between La Niña events and negative anomalies in southern Europe, especially in the southwest of the Iberian peninsula; however, they used a reduced number of stations. In this section we apply a simple statistical test to assess the teleconnection between ENSO and precipitation anomalies in Spain using the 123 gauges. This simple statistical methodology shall be extended later to validate seasonal forecasts.

[17] First, we computed the terciles for each station considering the series 1958–2004. This leads to three equiprobable categories: dry, normal and wet, each with a proportion $p_i = 1/3$ of the years. Then, the most extreme winter El Niño years (1958, 1966, 1969, 1973, 1983, 1987, 1988, 1992, 1995, 1998) and La Niña years (1965, 1971, 1974, 1976, 1985, 1989, 1999, 2000) were selected from the oceanic El Niño index provided by the Climatic Prediction Center (http://www.cpc.noaa.gov); note that the above years indicate the end of the corresponding ENSO events. These events are also in agreement with those found by *Pozo-Vázquez et al.* [2005] using a threshold based on the amplitude of the sea surface temperature normalized series over El Niño 3 region. For El Niño and La Niña periods, we obtained the proportions q_i (i = 1, 2, 3) of the dry/normal/wet categories in the different gauges, respectively, and tested whether the differences with the climatological proportions $p_i = 1/3$ were significant. To this aim we used a two-sided hypothesis test for the difference of two proportions [see, e.g., *Hahn and Meeker*, 1991] and obtained the associated significance level, or p-value p, of the test; note that usually the value $100 \times (1 - p)$% is used as the confidence level of the test (small p-values or, equivalently, high confidence values indicate high statistical significance of the result). Hence, we can test the relationship between the precipitation anomalies and El Niño (La Niña) events with a prescribed confidence level, allowing to determine whether an observed relationship (high or low proportion/probability of wet/normal/dry years) could simply occur by chance.

[18] Figure 3 shows the proportions (relative frequencies) and level of confidence for the dry/normal/wet categories during El Niño and La Niña events. The test is applied both to the raw data and to the gridded JRC observations. The first four rows correspond to El Niño and the last four to La Niña. Figure 3 shows a significant positive anomaly of dry days, especially in the middle North of the peninsula, for La Niña periods; over that area, the significance for some stations is very high. For El Niño events the relationship is much weaker and it is only significant at 90% confidence level in small isolated regions.

[19] Therefore, we could expect some conditional skill for seasonal forecast associated with La Niña events in some regions of Spain.

3. Statistical Downscaling

[20] In order to assess the skill of System2 seasonal precipitation forecasts we consider both the direct model outputs and the calibrated values obtained using a statistical downscaling technique. The statistical downscaling method considered in this paper is based on analogs, in particular a clustering weather typing variant introduced by *Gutiérrez et al.* [2004, 2005]. The method differs from the standard analog approach in the application of a clustering technique to the predictors. It allows to define meaningful subgroups (weather types) within the available circulation reanalysis. The subgroups are automatically defined using the ERA-40 reanalysis from the ECMWF from 1957 to 2002 [*Uppala et al.*, 2005], considering the geopotential height, temperature, specific humidity and eastward and northward wind components at 850 and 500 hPa in a temporal window covering the forecast day. In particular for day D we consider the temporal pattern given by the above variables at 00UTC for D and $D + 1$ (0 and 24 H). The temporal component of the patterns is included to cover the forecast period with all the available dynamical information (see *Gutiérrez et al.* [2004] for more details). Then, each System2 forecast is formed by

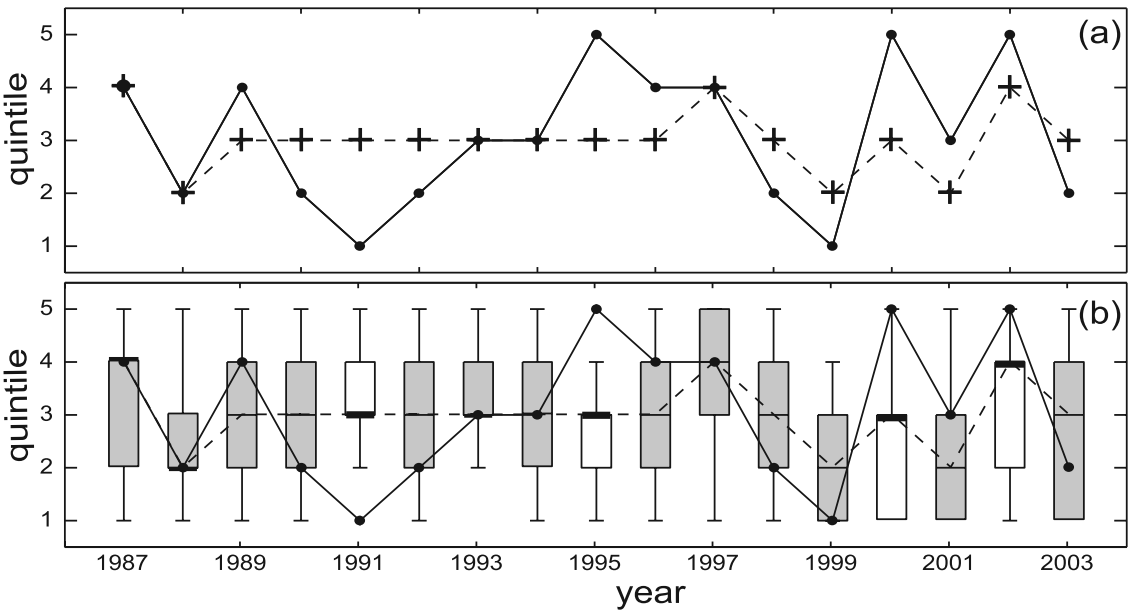

Figure 4. (a) Observed quintile (solid line) and predicted quintile by the ensemble median (dashed line) for precipitation data in a station over Spain. (b) Box plots of the quintiles corresponding to the members of the ensemble (the prediction interval is defined using the 25–75 percentiles, i.e., the boxes; whiskers represent the maximum and minimum values). Thick lines above or below the boxes indicate that the median coincides with the 75th or 25th percentiles, respectively. Moreover, shades indicate intervals capturing the observation.

40 members each providing 90 daily predictions for the DJF season. As shown by *Gutiérrez et al.* [2005], these patterns define a probabilistic distribution function (PDF) in the cluster space which can be combined with the median precipitation of each cluster to obtain the local seasonal downscaled precipitation for each member. We want to remark that although the train and validation periods overlap, a cross-validation study performed by removing the observations for a 1-year window around the analyzed season produced the same results; this is not surprising since seasonal forecasts are not expected to have a correspondence day by day with real observations.

[21] The regions used to define the predictor areas for the downscaling algorithm are the same used to aggregate the observed data (see Figure 1a).

4. Interval-Based Forecast

[22] The simplest approach to obtain a deterministic seasonal forecast from the ensemble is extracting a single value, such as the median, from the 40 members (note that each member gives a seasonal forecast value resulting from the 90 daily predictions). However this type of prediction does not take into account either the model systematic errors nor the spread of the forecast. To overcome the problem of systematic errors, instead of using standard bias removing procedures, we consider order statistics and, thus, the predictions (17 seasons $\times$ 40 members = 680 values) and observations (17 values) are divided separately into five equiprobable quintiles with its corresponding respective bounds (this procedure is similar to the cumulative distribution function matching method [e.g., see *Anagnostou et al.*, 1999]). This means that the forecast quantity is the quintile rather than precipitation per se. Note that using a discretization of the predictions in terms of quintiles is a reasonable first choice for seasonal forecast validation; for instance, this allows separating strong and moderate El Niño events, as shown by *Gutiérrez et al.* [2005], and it also provides a natural manner to eliminate the problem of systematic errors.

[23] The resulting forecasts can be simply validated applying a hit/miss procedure, comparing whether the predicted quintile (given by, e.g., the median of the member's predictions) coincides with the observed one (in the following the quintiles are denoted by the numbers 1, . . ., 5, to indicate the quantile intervals [0, 0.2], . . ., (0.8, 1], respectively). This idea is illustrated in Figure 4a and has the advantage of allowing simple inference tests to check the confidence on the obtained results, compared against a random forecast. For instance, a random forecast has an expected hit rate (HR, proportion of event occurrences that were correctly predicted, i.e., $P(predicted|occurred)$) of $p = 0.2$ (one out of five different possibilities). Moreover, an interval with 95% probability for the sampled HR can be easily obtained from the standard inference result for proportions: $p \pm z_\alpha\sqrt{(p-p^2)/n}$, where n is the number of years validated and z_α is given by the probability level. In this case, considering $n = 17$ and 95% of the probability, the interval (0.01, 0.39) is obtained. Therefore, a value of the HR larger than 0.2 but smaller than 0.39 indicates a performance of the method which is not significantly different from that of a random forecast and, hence, could just occur by chance with a confidence of 95%.

[24] To overcome the second problem (spread of the forecast), we follow the approach of *Weisheimer et al.* [2005] and consider a prediction interval spanning the

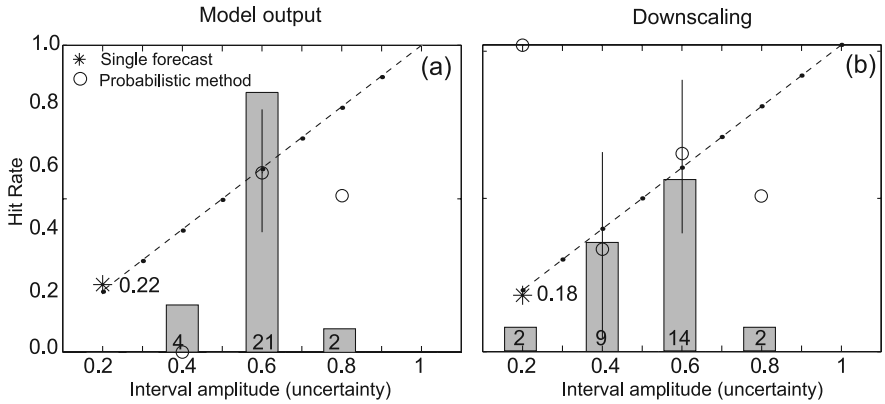

Figure 5. Joint verification for the two stations in Peru. The histogram shows the number of forecasts within each uncertainty level (numbers on the bottom). Hit rates (HR) corresponding to the ensemble median (asterisk) and to the interval-based method for different uncertainties (circles) are shown. Vertical lines represent the 90% confidence interval for those cases with more than six predictions. The expected HR for a random model is plotted by a dashed line. Results are presented from (a) the direct output from System2 and (b) the downscaling method.

uncertainty of the ensemble prediction system. However, instead of considering the minimum and maximum member predictions, we use a robust interval defined by the lower and upper quartiles (the interquartile range) of the 40 ensemble predictions. Then, the resulting prediction is not just a single quintile, but a subinterval of the quintile space [1,5] given by the lower and upper quartiles of the ensemble quintile-transformed predictions. For instance, Figure 4b illustrates this interval prediction approach for a particular station in Spain. The box-and-whiskers plots represent the empirical distributions of the quintiles predicted by the 40 ensemble members. Thus, the boxes correspond to the robust prediction intervals used in this study, which may capture (hit) or not (miss) the observed values (represented by dots), whereas the whiskers correspond to the bounding boxes, given by the minimum and maximum member predictions. Note that in this midlatitude region the bounding box covers the whole climatological interval in most of the cases and, thus, becomes useless as a prediction interval.

[25] This approach is applied to assess the performance of System2 seasonal precipitation predictions taking into account the uncertainty associated with the prediction, and considering both the direct model outputs and the values downscaled using the statistical technique described in section 3. Note that the model uncertainty is minimum when the interval reduces to a single quintile; in this case at least 50% of the members agree on the predicted quintile and the relative amplitude of the interval (defines as the fraction of quintiles spanned by the interval) reduces to 1/5. On the other hand, the model uncertainty is maximum when the predicted interval covers the whole quintile space [1, 5] with relative amplitude 5/5 = 1; in this case the prediction is useless since any quintile is probable to occur.

5. Verification Results

[26] The advantages of the proposed statistical approach to verify seasonal forecasts is first addressed over Peru to compare the results with those from previous studies, thus validating the methodology. Here, we consider the two stations analyzed by *Gutiérrez et al.* [2005, Figure 1]. Station data for two nearby sites located in the North of Peru (Morropón and Sausal de Culucán) were provided by the Meteorological Services of Peru (SENAMHI) for the period 1979–2001, but there is one missing season in the common period with System2 data (1987–2001). As shown by *Gutiérrez et al.* [2005] (see their Figure 2 for more details), precipitation in this tropical region is clearly influenced by the ENSO events. In particular, the data set from Morropón is more sensitive to this phenomenon. They found that heavy precipitation associated with strong El Niño events can be predicted some months in advance using a seasonal ensemble forecasts from the DEMETER project. The interval-based method is tested first over this tropical area to provide a more detailed information about predictability which includes the model uncertainty. The current study also gives an operational character to the analysis by using the System2 forecast system instead of the seasonal forecasts from the DEMETER project.

5.1. Interval-Based Validation of System2 Forecasts

[27] The hit rate (HR) is used in this part of the study to verify the precipitation predictions against the observations. First we computed the HR values corresponding to the deterministic prediction given by the quintile corresponding to the ensemble median precipitation, showing no skillful significant prediction (0.22 for Peru). The same conclusion was obtained for the downscaled precipitation (0.18). These results indicate that seasonal forecast does not provide skillful average results for the whole analysis period; however, there may be certain temporal windows of opportunity with skillful predictions. For instance, in the following we perform a validation conditioned to particular periods (for instance, associated with ENSO events) to look for seasonal forecast skill. For this purpose, we classified the predictions according to their uncertainty and validated each of the groups separately (note that predictions with low uncertainty are expected to have a larger skill). We used the relative amplitude of the predicted intervals as an indicator of the model uncertainty.

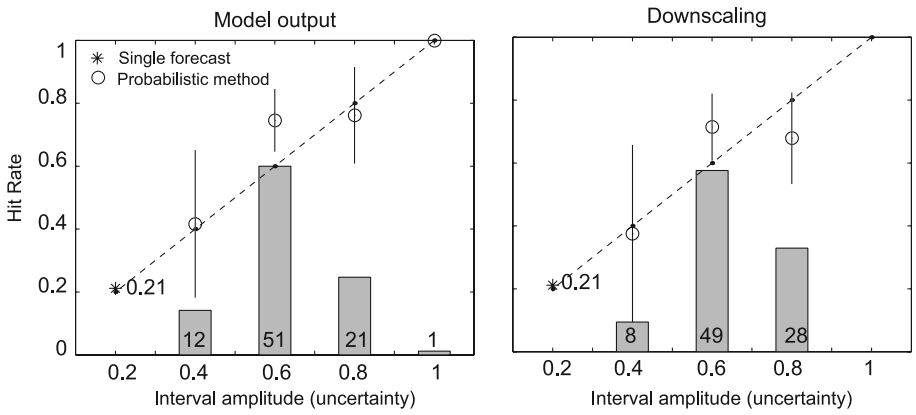

Figure 6. As in Figure 5, but for the five regions in Spain.

[28] Figure 5 shows the results for precipitation in Peru considering the two stations described above. Figure 5a represents the conditional validation for the direct System2 output, whereas Figure 5b shows the results when the downscaling method is applied to the seasonal system. The histogram represents the number of forecasts (numbers on the bottom) with different amplitudes, from a minimum value 0.2 (the interval reduces to a single quintile) to the maximum 1 (covering the whole climatological range). Note that a total of 14 (seasons) $\times$ 2 (stations) = 28 predictions are possible in this case, but there is one missing season in one of the stations; thus a total of 27 predictions are shown in Figure 5. Figure 5 shows that only a few System2 direct seasonal forecasts have low uncertainty, and this number is increased when the downscaling method is applied. However, a total of 11 predictions have amplitude smaller than 0.4 (i.e., the prediction interval covers at most two quintiles). The HR for different categories $p_i = P(\textit{predicted } q_i | \textit{occurred } q_i)$, $i = 1, \ldots, 5$, where q_i stands for the ith quintile, are represented by circles over the histogram bars and the 90% confidence intervals computed as $p_i \pm z_\alpha \sqrt{(p_i - p_i^2)/n_i}$, where n_i is the number of years falling in the ith category for the station under study, are plotted with vertical lines. The HR expected for a random forecast is indicated with a dashed line (note that the HR obtained with a random model increases linearly with the amplitude owing to the interval size). Here the random model is considered as the population, therefore the corresponding values are taken into account to evaluate the significant skill of the forecasts. The results corresponding to the downscaling show a significant skill (proportion significantly different from the random forecast) for the predictions with the smallest amplitude. In particular, the two forecasts included in this class correspond to the two predictions (one for each of the stations considered in Peru) for the strong 1997–1998 El Niño event. Figure 5 shows that the downscaling process is necessary in order to reduce the uncertainty of the prediction and gain skill (similar results have been also found by *Gutiérrez et al.* [2005], who also report the lack of local skill of the raw System2 data in this area).

[29] A similar study was performed in Spain, considering the five regions defined in section 2. Figure 6 shows the results for the System2 direct output and for the downscaled precipitation. In this case, predictions with significant skill correspond to the category of middle amplitude value (0.6), and there is no significant improvement of the results applying the downscaling method. Note that the predictions with this amplitude correspond to the intervals covering three quintiles: [1, 3], [2, 4], [3, 5] and indicate not wet, not

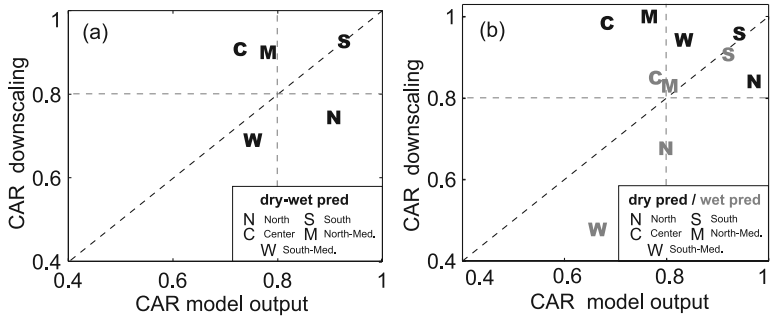

Figure 7. Correct Alarm Ratio (CAR) for the direct output versus the downscaled precipitation values for the five regions defined in Spain. (a) Combination of the dry and wet predictions, and (b) dry and wet predictions shown separately. For the sake of illustration, the vertical and horizontal dashed lines correspond to the upper bound of the 90% confidence interval for the CAR of a random forecast based on the $n = 17$ years.

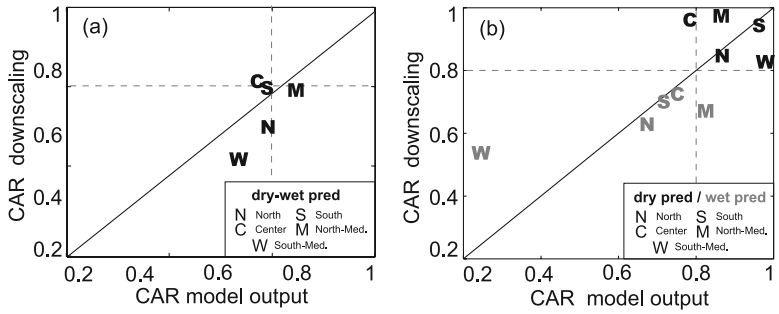

Figure 8. As in Figure 7 but using the gridded data from the JRC (Interpolated meteorological data Source JRC/AGRIFISH Data Base-EC-JRC).

extreme wet or dry, and not dry, respectively. Therefore, this category of predictions indicates the no occurrence of different events, instead the occurrence of events and this corresponds to weaker signals. Thus, in this case the seasonal forecast system seems to be skillful only to discard the occurrence of a given event. Similar results (not shown) were obtained when the individual stations were validated separately instead of using regional aggregation.

5.2. Validation Conditioning on the Prediction Values

[30] In this section we analyze in more detail the predictability found in Spain associated with middle uncertainty intervals (those with amplitude 0.6). In order to identify the source of predictability, only the forecasts with extreme values (including quintiles 1 or 5) were considered; therefore, we pay attention to the highest fluctuations of the model, discarding normal or intermediate predictions. Note that in this case we need to condition the validation on the predictions, and not on the occurrences. Therefore, instead of considering the HR, i.e., $P(predicted|occurred)$, we compute the "Correct-Alarm Ratio" (CAR) [*Mason and Graham*, 1999], given by $P(occurred|predicted)$, i.e., the proportion of correct predicted events. This index considers the skill from the perspective of the forecasts not from the observations. This information is valuable for the end users because it provides an indicator about how likely is the occurrence of an event given that the "test" (the prediction in this case) is positive. Figure 7 compares the values of the CAR for the downscaled precipitation and the direct System2 output for the five regions over Spain. Figure 7a depicts the CAR for the wet and dry periods; Figure 7b shows separately negative and positive anomalies using different colors, black and grey respectively. The main result from Figure 7 is that the downscaling method displays high CAR values corresponding to negative anomalies, indicating predictability in boreal winter over Spain related to drought events. We checked the years of the predictions with higher skill (1989 and 2000 episodes in most of the regions) obtaining an agreement with the teleconnection with La Niña analyzed in section 2.3. Note that the regions with skill do not exactly match the ones with significant teleconnection observed in Figure 3; however, a shorter period of time is available (1987–2001). The main conclusion of this part is that operational seasonal forecasting systems are able to reproduce the observed teleconnection described in section 2.3, since they show skill to predict the wet or dry events in Spain associated with ENSO anomalies. Thus, it is shown that the known predictability of the tropical ocean in combination with the existing teleconnections with midlatitude regions provide a window of opportunity for seasonal forecasting in Spain. A more general study may lead to further advances of seasonal forecast skill in Europe, where ENSO teleconnections have been also reported [*van Oldenborgh et al.*, 2000].

6. Sensitivity Studies

[31] The results presented in the previous section assess the conditional skill of System2 to predict boreal winter precipitation over Peru and Spain, and report the improvements obtained using an analog downscaling method. However, there are some factors which can exert some influence on those results, such as the nature of the observations (raw or interpolated) used to validate the forecasts and to fit the downscaling method, or the temporal aggregation (daily, weekly or monthly) of the atmospheric patterns considered in the downscaling method. For this reason, it is necessary to analyze the sensitivity of the results to those aspects, in order to assess the consistency of the results. These characteristics are discussed in the next sections. We will consider only the case of Spain, where the results seem to be weaker than in Peru.

6.1. Raw Versus Interpolated Observations

[32] The quality of the observed data set considered for the target variable can influence the results derived from the study. Station data have been used in section 5 to validate the predictions from System2 and to train and validate the downscaling method. However, the use of gridded high-resolution data derived from observations has the advantage of correcting inhomogeneities and possible errors in the observations. Nevertheless, some information from the original local data can be left outside during that process which will probably modify in some extent the results. The consequences of this change applied to the original information are assessed in this section focusing the study on the variations of the seasonal forecast system skill.

[33] Figure 8 shows the CAR values for the predicted wet and dry precipitation taking into account the gridded data set from the JRC (see section 2). The comparison of the results shown in Figures 7 and 8 shows that now both the model output and the downscaling exhibit skill to predict

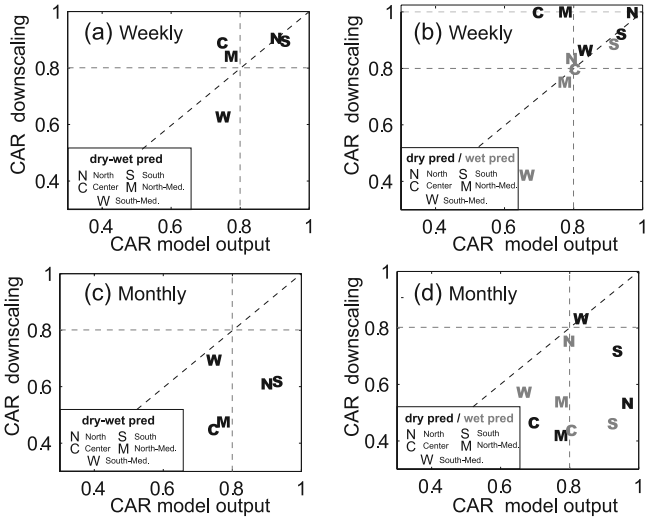

Figure 9. As in Figure 7 but considering (top) weekly and (bottom) monthly atmospheric patters in the downscaling process.

dry events. This is not a surprising result because the gridded observations are more suitable for raw model output validation, since they are also representative of a grid-box area instead of a local value.

6.2. Temporal Scale for the Downscaling Method

[34] One important factor influencing the downscaling method is the choice of the temporal aggregation scale used to define the atmospheric patterns (daily for the previous results in section 5). This aspect plays a relevant role on the downscaling process since the resulting weather types depend on that selection. The use of daily patterns involves a high computational and storage cost that might not be justified. At this point we address a simple question. What is the optimum temporal scale to define the atmospheric patterns at which the downscaling method is able to account for most information with low computational cost? Since the main goal of the study is focused on seasonal scale it can be more reasonable to take into account weekly or even monthly atmospheric patterns, since these temporal scales are less computationally expensive.

[35] The CAR was calculated again for the dry and wet events using station data, but considering weekly and monthly atmospheric patterns, respectively. Figure 9 shows the values using weekly patterns (top) and monthly patterns (bottom). From these results it is clear that, on the one hand, the use of monthly atmospheric patterns lead to downscaled predictions with no skill. Thus, this temporal scale is not appropriate for the proposed downscaling method since relevant circulation information is missing from the pattern. On the other hand, the use of weekly patterns outperforms the results obtained with daily values (Figure 7) and also provide a better match with the spatial pattern of the teleconnections found in section 2.3 (associated with the north and central regions). Therefore, weekly patterns seem to provide an appropriate temporal resolution for downscaling seasonal forecasts in Spain. This result is in agreement with previous studies [*Zorita and von Storch*, 1999], but further research is necessary to better understand the effect of the temporal aggregation in the seasonal predictability.

7. Conclusions

[36] The interval-based validation method presented in this paper is able to evaluate the skill of a seasonal prediction ensemble system, providing also an estimation of the statistical significance. Using this approach seasonal forecasts from System2 is shown to have high predictability for boreal winter precipitation over Peru during El Niño episodes, in accordance with previous studies. These events correspond to forecasts with the lowest uncertainty. Over this region, the need of regional information has been proved and the downscaling approach clearly improves the forecasting system skill. To properly assess this result it would be necessary to analyze the predictability taking into account more strong El Niño events by using longer data sets, for instance using the DEMETER outputs, although in this case we would dismiss the interest of using an operational forecasting system.

[37] As expected, the predictability is lower at higher latitudes where in general the ensemble spread is higher. However, the interval-based method is able to uncover some winter precipitation predictability over Spain related to drought episodes. This fact is explained by a known teleconnection between these negative extreme episodes and La Niña events. As we show, this teleconnection is reproduced by the seasonal forecast system and provides a window of opportunity for operational seasonal forecast in midlatitudes.

[38] Some preliminary sensitivity studies indicate that a higher skill is obtained for the direct model outputs when using gridded observations instead of raw data in local stations for validation; this difference is not observed when validating the statistical downscaled values (since the downscaling process adapts/callibrate the outputs to the observed data). It is also shown that the weekly scale for the atmospheric patterns seems to be the optimum temporal scale for seasonal forecast in Spain. At that timescale the downscaling method is able to account for most information with low computational cost.

[39] Finally, note that the seasonal forecasts can be validated at any desired resolution (terciles, quintiles, deciles), keeping a balance with the available data to allow the application of significance tests.

[40] **Acknowledgments.** The authors want to thank the anonymous referees for their comments and criticisms, which helped us to improve the paper. The authors are also grateful to the Comisión Interministerial de Ciencia y Tecnología (CICYT, CGL2004-02652 and CGL2005-06966-C07-02/CLI grants) for partial support of this work. The authors also acknowledge the Spanish State Meteorological Agency (AEMET) and Joint Research Centre–European Commission (JRC) for the raw and interpolated meteorological data sources provided for this work, respectively.

References

Anagnostou, E. N., A. J. Negri, and R. F. Adler (1999), Statistical adjustment of satellite microwave monthly rainfall estimates over Amazonia, *J. Appl. Meteorol.*, *38*, 1590–1598.

Anderson, D., T. Stockdale, L. Ferranti, and M. Balmaseda (2003), The ECMWF seasonal forecasting system, *ECMWF Newsl.*, *98*, 17–25.

Balakrishnan, N., and A. C. Cohen (1991), *Order Statistics and Inference: Estimation Methods*, Academic Press, San Diego, Calif.

Challinor, A. J., J. M. Slingo, T. R. Wheeler, and F. J. Doblas-Reyes (2005), Probabilistic simulations of crop yield over western India using the DEMETER seasonal hindcast ensembles, *Tellus, Ser. A*, *57*, 498–512.

Derome, J., H. Lin, and G. Brunet (2005), Seasonal forecasting with a simple general circulation model: Predictive skill in the AO and PNA, *J. Clim.*, *18*, 597–609.

Díez, E., C. Primo, J. A. García-Moya, J. M. Gutiérrez, and B. Orfila (2005), Statistical and dynamical downscaling of precipitation over Spain from DEMETER seasonal forecasts, *Tellus, Ser. A*, *57*, 409–423.

Douville, H. (2004), Relevance of soil moisture for seasonal atmospheric predictions: Is it an initial value problem?, *Clim. Dyn.*, *22*, 429–446.

Gershunov, A., and D. R. Cayan (2003), Heavy daily precipitation frequency over the contiguous United States: Sources of climatic variability and seasonal predictability, *J. Clim.*, *16*, 2752–2765.

Gutiérrez, J. M., A. S. Cofiño, R. Cano, and M. A. Rodríguez (2004), Clustering methods for statistical downscaling in short-range weather forecasts, *Mon. Weather Rev.*, *132*, 2169–2183.

Gutiérrez, J. M., R. Cano, A. S. Cofiño, and C. Sordo (2005), Analysis and downscaling multi-model seasonal forecasts in Peru using self-organizing maps, *Tellus, Ser. A*, *57*, 435–447.

Hahn, G. J., and W. W. Meeker (1991), *Statistical Intervals: A Guide for Practitioners*, Wiley-Interscience, Hoboken, N. J.

Hastenrath, S. (1995), Recent advances in tropical climate prediction, *J. Clim.*, *67*, 1519–1532.

Johansson, A., A. Barnston, S. Saha, and H. van den Dool (1998), On the level and origin of seasonal forecast skill in northern Europe, *J. Atmos. Sci.*, *55*(1), 103–127.

Jolliffe, I. T., and D. B. Stephenson (2003), *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, John Wiley, Hoboken, N. J.

Mason, S., and N. Graham (1999), Conditional probabilities, relative operating characteristics and relative operating levels, *Weather Forecasting*, *14*, 713–725.

Muñoz-Díaz, D., and F. Rodrigo (2004), Spatio-temporal patterns of seasonal rainfall in Spain (1912–2000) using cluster and principal component analysis: Comparison, *Ann. Geophys.*, *22*, 1435–1448.

Osborn, T. J., and M. Hulme (1997), Development of a relationship between station and grid-box rainday frequencies for climate model evaluation, *J. Clim.*, *10*, 1885–1908.

Palmer, T. N., and D. L. T. Anderson (1994), The prospects for seasonal forecasting—A review paper, *Q. J. R. Meteorol. Soc.*, *120*, 755–793.

Palmer, T. N., et al. (2004), Development of a European multimodel ensemble system for seasonal-to-interannual prediction DEMETER, *Bull. Am. Meteorol. Soc.*, *85*, 853–872.

Pozo-Vázquez, D., S. R. Gámiz-Fortis, J. Tovar-Pescador, M. J. Esteban-Parra, and Y. Castro-Díez (2005), El Niño–Southern Oscillation events and associated European winter precipitation anomalies, *Int. J. Climatol.*, *25*, 17–31.

Quan, X. W., M. P. Hoerling, J. S. Whitaker, G. T. Bates, and T. Y. Xu (2006), Diagnosing sources of U.S. seasonal forecast skill, *J. Clim.*, *19*, 3279–3293.

Saha, S., et al. (2006), The NCEP climate forecast system, *J. Clim.*, *19*, 3483–3517.

Thompson, M. C., F. J. Doblas-Reyes, S. J. Mason, R. Hagedorn, S. J. Connor, T. Phindela, A. P. Morse, and T. N. Palmer (2006), Malaria early warnings based on seasonal climate forecasts from multi-model ensembles, *Nature*, *439*, 576–579.

Uppala, S. M., et al. (2005), The ERA-40 re-analysis, *Q. J. R. Meteorol. Soc.*, *131*, 2961–3012.

van Oldenborgh, G. J., G. Burgers, and A. K. Tank (2000), On the El Niño teleconnection to spring precipitation in Europe, *Int. J. Climatol.*, *20*, 565–574.

Wang, G., R. Kleeman, N. Smith, and F. Tseitkin (2001), The BMRC coupled general circulation model ENSO forecast system, *Mon. Weather Rev.*, *130*, 975–991.

Wang, W., and A. Kumar (1998), A GCM assessment of atmospheric seasonal predictability associated with soil moisture anomalies over North America, *J. Geophys. Res.*, *103*, 637–646.

Weisheimer, A., L. A. Smith, and K. Judd (2005), A new view of seasonal forecast skill: Bounding boxes from the DEMETER ensemble forecasts, *Tellus, Ser. A*, *57*, 265–279.

Weiss, E. B. (1982), The value of seasonal climate forecasts in managing energy resources, *J. Appl. Meteorol.*, *21*, 510–517.

Zorita, E., and H. von Storch (1999), The analog method as a simple statistical downscaling technique: Comparison with more complicated methods, *J. Clim.*, *12*, 2474–2489.

A. S. Cofiño, M. D. Frías, J. M. Gutiérrez, S. Herrera, and C. Sordo, Department of Applied Mathematics and Computer Science, University of Cantabria, Avenida de los Castros, E-39005 Santander, Spain. (friasmd@unican.es)